

1 DISTRIBUTED RESOURCE RESERVATION SYSTEM FOR ESTABLISHING A
2 PATH THROUGH A MULTI-DIMENSIONAL COMPUTER NETWORK TO SUPPORT
3 ISOCHRONOUS DATA
4

5 CROSS REFERENCE TO RELATED APPLICATIONS AND PATENTS

6 This application is related to co-pending patent application serial no. 09/542,954 entitled
7 "METHOD FOR DESIGNATING ONE OF A PLURALITY OF ADDRESSABLE STORAGE
8 DEVICES TO PROCESS A DATA TRANSFER REQUEST" filed on 3/31/2000, the disclosure
9 of which is incorporated herein by reference.
10

11 BACKGROUND OF THE INVENTION

12
13 Field of the Invention

14 The present invention relates to computer networks. More particularly, the present
15 invention relates to a distributed resource reservation system for establishing a path through a
16 multi-dimensional computer network to support isochronous data.

17 Description of the Prior Art

18 Establishing a path through a network to handle isochronous traffic has in the past been
19 managed by a central routing processor which evaluates the state of the network and reserves
20 resources in a path from a destination to source node. The resources are reserved in a manner
21 that guarantees a delivery time for the isochronous data. However, the central processor may
22 require a significant amount of memory and processing power to evaluate the network and
23 reserve resources to establish paths in an efficient and timely manner. Further, as the network
24 grows, so must the memory and processing capabilities of the central server. Thus, centrally
25 managed schemes do not scale well.

26 The Resource Reservation Protocol or RSVP protocol is a prior art method for reserving
27 resources to support isochronous traffic in a network wherein the resource reservation facility is

distributed throughout the nodes of the network (see the Resource Reservation Protocol or RSVP an overview for which is provided in the text book *Managing Bandwidth - Deploying QOS in Enterprise Networks*, by Alistair Croll and Eric Packman, Prentice Hall, Upper Saddle River, NJ, 1999; and "Resource Reservation Protocol (RSVP) -- Version 1 Functional Specification", Braden, R., Zhang, L., Berson, S., Herzog, S., Jamin, S., RFC 2205, September 1997, Proposed Standard).. The RSVP protocol is illustrated in FIG. 1A which shows a prior art tree network comprising a plurality of nodes wherein a destination node reserves resources in a path toward a source node. In FIG. 1A, node 4A sends a request packet into the network 2 to receive isochronous data from node 4D. As the request packet travels through the network 2, each node in the path (e.g., node 4B and 4C) determines whether it has sufficient resources to service the request and then reserves the resources. FIG. 1B shows a scenario when node 4C cannot service a request from node 4A due to resources having already been reserved to service a request from node 4E. Node 4C sends a reply message to node 4A indicating that the request cannot be serviced. Node 4A must then wait until node 4C has regained sufficient resources to service the request (e.g., when node 4E relinquishes the resources reserved in node 4C). This blocking characteristic of busses and tree networks implementing the RSVP protocol can reduce the aggregate performance of the network, particularly during high traffic periods.

There is, therefore, a need to establish paths for isochronous data in a computer network in a cost effective manner which scales efficiently with the size of the computer network and which avoids the blocking characteristic of the RSVP protocol.

SUMMARY OF THE INVENTION

The present invention may be regarded as a distributed method of establishing a path in a multi-dimensional computer network comprising a plurality of nodes for transmitting isochronous data from a source node to a destination node. A request packet is injected into the network, the request packet specifying a request to transmit the isochronous data from one of a plurality of source nodes. The request packet is routed to at least one of the plurality of source nodes, and the source node determines whether it has sufficient resources to support transmitting

1 the isochronous data. If the source node comprises sufficient resources to support transmitting
2 the isochronous data, the source node reserves resources within the source node to support
3 transmitting the isochronous data, and transmits an acknowledge (ack) packet from the source
4 node to a first neighboring node. The first neighboring node determines whether it has sufficient
5 resources to support transmitting the isochronous data. If the first neighboring node comprises
6 sufficient resources to support transmitting the isochronous data, the first neighboring node
7 reserves resources within the first neighboring node to support transmitting the isochronous data,
8 and transmits the ack packet to an adjacent node. If the first neighboring node does not comprise
9 sufficient resources to support transmitting the isochronous data, the first neighboring node
10 transmits a negative-acknowledge (nack) packet to the source node and the source node transmits
11 the ack packet to a second neighboring node.

12 In one embodiment, the resources reserved within a node comprises memory for
13 buffering data, and in another embodiment the resources comprises network communication
14 circuitry, such as a virtual lane in switching circuitry.

15 In another embodiment, the resources are reserved within a node for a specified "lease"
16 period, after which time the resources are automatically relinquished. In yet another
17 embodiment, the resources are automatically relinquished if they remain idle for a specified
18 "time-out" period.

19 In still another embodiment, the request is routed from the destination node to a plurality
20 of source nodes and a method is employed to determine which source node will service the
21 request. In one embodiment, the request packet is transmitted to a primary source node, the
22 primary source node determines whether it has sufficient resources to service the request and, if
23 not, the primary source node transmits the request packet to a secondary source node. In another
24 embodiment, the request packet is multicast to a plurality of source nodes, at least two of the
25 source nodes reserves resources to support transmitting the isochronous data, and one of the
26 reservations is canceled when a path constructed from the source node encounters a path which
27 has already reserved resources to support transmitting the isochronous data.

The present invention may also be regarded as a switched node comprising switching circuitry including more than two bi-directional ports for simultaneously transmitting data in multiple dimensions through a computer network, wherein each bi-directional port comprises an input port and an output port. The switched node further comprises a data buffer for buffering data, routing circuitry for routing data stored in the data buffer to a selected output port, and a reservation facility for reserving resources within the switch node to support requests to transmit isochronous data. The switched node receives a request packet to reserve resources to support transmitting isochronous data. If the switched node comprises sufficient resources to support transmitting the isochronous data, the reservation facility reserves resources within the switched node to support transmitting the isochronous data, and the switched node transmits an ack packet to a first neighboring node. If the first neighboring node does not comprise sufficient resources to support transmitting the isochronous data, the switched node receives a nack packet from the first neighboring node, and the switched node transmits the ack packet to a second neighboring node.

In one embodiment, the switched node further comprises a disk for storing data and a head actuated over the disk for writing data to and reading data from the disk. In one embodiment, the reservation facility reserves resources associated with data read from the disk and written to the disk.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1A illustrates the prior art resource reservation protocol (RSVP) wherein each node in a network tree comprises reservation facilities for establishing isochronous data paths.

FIG. 1B illustrates how in the RSVP protocol a previously established path can block requests to create new paths.

FIG. 2A illustrates a distributed resource reservation system according to an embodiment of the present invention wherein a request to transmit isochronous data is sent to a source node in a multi-dimensional computer network.

FIG. 2B illustrates how the source node of FIG. 2A generates an acknowledge (ack)

1 packet which is transmitted back to the destination node along a path wherein resources are
2 reserved at each node to support transmission of the isochronous data.

3 FIG. 2C shows how a path is redirected around a node which does not have sufficient
4 resources to handle the isochronous data.

5 FIG. 2D shows how the path established in FIG. 2B is torn down by transmitting a
6 termination packet through the path from the source node to the destination node.

7 FIG. 3A shows an embodiment of the present invention wherein the request to receive the
8 isochronous data is forwarded to a secondary source node if a primary source node does not
9 comprise sufficient resources to service the request.

10 FIG. 3B shows the path established by the secondary source node of FIG. 3A used to
11 transmit the isochronous data.

12 FIG. 4A shows an embodiment of the present invention wherein the request to receive the
13 isochronous data is multicast to a plurality of nodes in the multi-dimensional computer network.

14 FIG. 4B shows two source nodes reserving resources through two paths toward the
15 destination node to support transmitting of the isochronous data.

16 FIG. 4C illustrates how one of the paths in FIG. 4B is torn down when a node is
17 encountered which has already reserved resources to support transmitting the isochronous data
18 through an alternative path.

19 FIG. 5 shows a switched node for use in a multi-dimensional switched fabric computer
20 network according to an embodiment of the present invention comprising a plurality of bi-
21 directional ports for simultaneously transmitting data in multiple dimensions, a data buffer for
22 buffering data, routing circuitry for routing the data from the data buffer to the appropriate output
23 ports, and a reservation facility for reserving resources within the switch node to support requests
24 to transmit isochronous data.

25 FIG. 6 shows the switched node of FIG. 5, further comprising a disk for storing data and
26 a head actuated over the disk for writing data to and reading data from the disk, and a reservation
27 facility for reserving resources associated with data read from the disk and written to the disk.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

FIGs. 2A-2C illustrates a distributed method of establishing a path in a multi-dimensional computer network 6 comprising a plurality of nodes for transmitting isochronous data from a source node 8 to a destination node 10 according to an embodiment of the present invention. A request packet 12 is injected into the network, the request packet 12 specifying a request to transmit the isochronous data from one of a plurality of source nodes. The request packet is routed to at least one of the plurality of source nodes 8, and the source node 8 determines whether it has sufficient resources to support transmitting the isochronous data. If the source node 8 comprises sufficient resources to support transmitting the isochronous data, the source node 8 reserves resources within the source node 8 to support transmitting the isochronous data, and transmits an acknowledge (ack) packet 14 from the source node 8 to a first neighboring node 16 as shown in FIG. 2B. The first neighboring node 16 determines whether it has sufficient resources to support transmitting the isochronous data. If the first neighboring node 16 comprises sufficient resources to support transmitting the isochronous data, the first neighboring node 16 reserves resources within the first neighboring node 16 to support transmitting the isochronous data, and transmits the ack packet 14 to an adjacent node 18. As shown in FIG. 2C, if the first neighboring node 16 does not comprise sufficient resources to support transmitting the isochronous data, the first neighboring node 16 transmits a negative-acknowledge (nack) packet 20 to the source node 8 and the source node 8 transmits the ack packet 14 to a second neighboring node 22.

In one embodiment, the request packet 12 specifies certain Quality-of-Service (QOS) constraints associated with the isochronous data (e.g., transmission latency and data rate). The nodes within the network 6 determine whether they have sufficient resources to support the QOS constraints. Once a path 24 has been established between the source node 8 and the destination node 10, the reserved resources guarantees that the isochronous data is transmitted through the path 24 in a manner that satisfies the specified QOS constraints.

In the embodiment of FIG. 2A, the request packet 12 is injected into the network 6 by the

1 destination node 10. In alternative embodiment, the request packet 12 may be injected into the
2 network 6 by a third party node which requests that the isochronous data be delivered to the
3 destination node 10.

4 When the source node 8 finishes transmitting the isochronous data to the destination node
5 10, the path 24 reserved for the data transfer is torn down by relinquishing the resources reserved
6 at each node through the path. In one embodiment illustrated in FIG. 2D, the source node 8
7 transmits a termination packet 26 through the path 24 after transmitting the last packet associated
8 with the isochronous data. As the termination packet 26 reaches each node in the path 24, the
9 nodes relinquish the resources reserved for the isochronous data.

10 In another embodiment, the destination node specifies a "lease" period for the resources
11 reserved within the nodes, and the resources are automatically relinquished when the lease period
12 expires. In yet another embodiment, a time-out period is associated with the resources reserved
13 within the nodes and the resources are automatically relinquished if they remain idle beyond the
14 time-out period. This embodiment prevents blocking of resources under failure conditions, such
15 as the source node 8, the destination node 10, or an intermediate node within the path failing. In
16 yet another embodiment, the destination node 10 may intentionally terminate the transmission
17 early by sending a termination packet through the path which also relinquishes the resources
18 reserved for the isochronous data.

19 FIG. 3A shows an embodiment of the present invention wherein the request packet 12 is
20 first transmitted to a primary source node 8 which determines whether it comprises sufficient
21 resources to support transmitting the isochronous data. If the primary source node 8 does not
22 comprise sufficient resources, then the request packet 12 is forwarded to a secondary node 26
23 comprising a replicate of the data stored on the primary source node 8. If the secondary node 26
24 has sufficient resources, then it sends an ack packet 30 through nodes in the network to establish
25 a path 32 for the isochronous data. In one embodiment, the primary source node 8 stores
26 information identifying the secondary source node 26 to which the request packet 12 is
27 forwarded. In an alternative embodiment, the request packet 12 comprises information

1 identifying the secondary source node 26.

2 FIG. 4A illustrates an embodiment of the present invention wherein the request packet 12
3 is multicast to a plurality of source nodes in the network attempting to find at least one source
4 node comprising sufficient resources to support transmitting the isochronous data. In one
5 embodiment shown in FIG. 4B, resources are reserved in multiple paths through the network. In
6 the example of FIG. 4B, source nodes 8 and 26 both comprise sufficient resources to support
7 transmitting the isochronous data. Both source nodes 8 and 26 reserve resources and begin to
8 establish a first path and a second path toward the destination node 10 by transmitting a first ack
9 packet 28 and a second ack packet 30 through the network. When the first ack packet 28 reaches
10 a node 32 comprising resources already reserved to support transmitting the isochronous data
11 through the second path, the resources reserved in the first path are relinquished. In one
12 embodiment shown in FIG. 4C, the ack packet 28 is converted into a cancel packet 34 which is
13 transmitted back toward the source node 26 relinquishing the resources reserved at each node.
14 The isochronous data is then transmitted through the path 36 established by source node 8.

15 In an alternative embodiment, the request packet 12 is multicast to a plurality of source
16 nodes and a facility is provided to select from the multiple source nodes that comprise sufficient
17 resources to support transmitting the isochronous data. For example, the source nodes may
18 communicate with one another or with a central processor to determine the most appropriate
19 source node to service the request. The above reference co-pending patent application entitled
20 "METHOD FOR DESIGNATING ONE OF A PLURALITY OF ADDRESSABLE STORAGE
21 DEVICES TO PROCESS A DATA TRANSFER REQUEST " discloses further details of this
22 embodiment.

23 FIG. 5 shows a switched node 40 for use in a multi-dimensional computer network
24 according to an embodiment of the present invention. The switched node 40 comprises
25 switching circuitry comprising more than two bi-directional ports for simultaneously transmitting
26 data in multiple dimensions through the computer network, wherein each bi-directional port
27 comprises an input port 42A-42D and an output port 44A-44D, a data buffer 46A-46D for

1 buffering data, routing circuitry for routing data stored in the data buffer to a selected output port
2 44A-44D, and a reservation facility 48A-48D for reserving resources within the switch node 40
3 to support requests to transmit isochronous data. The switched node 40 receives a request
4 packet to reserve resources to support transmitting isochronous data. If the switched node 40
5 comprises sufficient resources to support transmitting the isochronous data, the reservation
6 facility 48A-48D reserves resources within the switched node 40 to support transmitting the
7 isochronous data, and the switched node 40 transmits an ack packet to a first neighboring node.
8 If the first neighboring node does not comprise sufficient resources to support transmitting the
9 isochronous data, the switched node 40 receives a nack packet from the first neighboring node,
10 and the switched node 40 transmits the ack packet to a second neighboring node.

11 The switched node 40 of FIG. 5 is interconnected with a plurality of other switched nodes
12 such as shown in FIG. 2A to form a multi-dimensional switched fabric. Each of the switched
13 nodes in FIG. 2A comprises four bi-directional ports (North, East, South and West) forming a
14 two-dimensional fabric. In one embodiment, the network data transmitted through the switched
15 nodes consist of packets having a packet header comprising routing data which identifies the
16 source node for the packet. The packet headers are processed in order to route the packet through
17 the switched nodes. A suitable routing algorithm generates control data for configuring the
18 switched nodes in order to route the packets through the network. Any suitable routing
19 algorithm may be employed, and it may support Unicast, Multicast, or Broadcast delivery
20 mechanisms. The routing decisions may be made centrally, at the source, distributed, or
21 multiphase, implemented using a lookup table or using a finite-state machine. Further, the
22 routing algorithm may be deterministic or adaptive. A discussion of various routing algorithms
23 which may be employed in the embodiments of the present invention is provided by Jose Duato
24 et al. in the textbook "Interconnection Networks, an Engineering Approach", IEEE Computer
25 Society, 1997.

26 The routing algorithm is implemented a layer "above" the switching layer, and thus the
27 routing algorithm may be compatible with various different switching algorithms, for example,

1 Virtual Cut-Through Switching, Wormhole Switching, and Mad Postman Switching. In
2 addition, topologies other than the two-dimensional switched fabric of FIG. 2A, as well as
3 topologies comprising more than two dimensions, may be employed in the present invention by
4 decreasing or increasing the number of bi-directional ports per switched node. Various
5 topologies and switching algorithms which may be employed in the embodiments of the present
6 invention are discussed in the aforementioned textbook by Jose Duato et al.

7 Referring again to FIG. 5, the bi-directional ports of the switched node 40 comprise four
8 input ports 42A-42D and four output ports 44A-44D corresponding to the North, East, South and
9 West ports shown in FIG. 2A. Data packets received from the input ports 42A-42D are buffered
10 in FIFO buffers 50A-50D. A routing table 52 is configured by control data 54 which in one
11 embodiment is generated by a central routing microprocessor. The routing table 52 generates
12 control signals 56A-56D which configure multiplexers 58A-58D in order to route the data
13 packets to appropriate data buffers 46A-46D associated with the output ports 44A-44D. In this
14 manner, the data packets cross the switched node 40 immediately except for the delay of the
15 FIFO buffer 50A-50D. The FIFO buffers 50A-50D provide buffering of input data in the event
16 that the target data buffer 46 is full or busy receiving data from another of the input ports.

17 A scheduling facility 60A-60D is also provided within the switched node 40 which
18 schedules the time when the data packets are to be transferred from the data buffers 46A-46D to
19 the output ports 44A-44D. In one embodiment, the timing information for the packets are stored
20 in the packet headers and processed by the scheduling facility 60A-60D. In one embodiment, the
21 timing information implements an isochronous communication protocol such as disclosed in the
22 in Texas Instruments' TSB12LV41A link-layer controller (LLC) which supports the IEEE 1394
23 specification for high-performance serial bus with automatic generation of the common
24 isochronous packet headers and time stamping as required by the IEC 61883 standard.

25 In one embodiment, the data buffers 46A-46D comprise a plurality of virtual lanes where
26 each virtual lane is assigned a predetermined priority level. The scheduling facility 60A-60D
27 schedules the data packets according to the timing information by queuing the data packets in the

appropriate virtual lanes. For example, data packets with shorter transmission deadlines are queued in higher priority virtual lanes, whereas data packets with longer transmission deadlines are queued in lower priority virtual lanes. In addition, within a virtual lane the data packets can be queued in order of arrival (FIFO) or in order of departure based on the transmission deadlines in order to support predetermined QOS constraints. Details of departure queuing are disclosed by Jennifer Rexford, et al. in "A Router Architecture for Real-Time Communication in Multicomputer Networks", *IEEE Transactions on Computers*, Vol. 47, No. 10, October 1998, which is incorporated herein by reference.

In one embodiment, the reservation facility 48A-48D reserves one or more of the virtual lanes for transmitting isochronous data. In another embodiment, the reservation facility 48A-48D reserves memory in data buffers 46A-46D for transmitting isochronous data. In yet another embodiment, the reservation facility 48A-48D reserves processing circuitry used to implement the routing and scheduling operations within the switched node 40. In still another embodiment, the reservation facility 48A-48D reserves bandwidth of linking circuitry for linking the switched node 40 to other switched nodes in the computer network.

The switched node of 40 of FIG. 5 can be extended to add additional dimensionality by duplicating the circuitry associated with each bi-directional port (input port 42, FIFO 50, MUX 58, output port 44, etc.). In one embodiment, the switched node 40 is a commodity device which comprises a facility for dynamically configuring the bi-directional ports to support a desired switched fabric topology. Thus, a number of the input ports 42A-42D and/or a number of the output ports 44A-44D may be configured to connect to ports of other switched nodes, whereas the remaining ports may be left unconnected.

FIG. 6 shows a switched node 62 for use in a multi-dimensional computer network according to an alternative embodiment of the present invention wherein a disk 64a and a head 64b are integrated into the switched node 62 for storing isochronous data. The data packets received from the input ports 42A-42D may be routed to an input port 66 associated with the disk 64a and the head 64b, where the data is stored in data buffer 68, and ultimately written onto the

1 disk 64a. Data read from the disk 64a is also stored in the data buffer 68 and transmitted via
2 output port 70 to the appropriate data buffer 46A-46D.

3 In the embodiment of FIG. 6, the reservation facility 48A-48D within the switched node
4 62 reserves resources associated with the scheduling facility 60A-60D to support transmitting
5 isochronous data through the switched node 62, and a reservation facility 48E reserves resources
6 associated with a scheduling facility 60E to support isochronous data transmitted to and from the
7 disk 64a. The scheduling facility 60E also comprises additional resources for implementing the
8 interface between the data buffer 68 and the disk 64a. In one embodiment, the reservation
9 facility 48E reserves memory within the data buffer 68 to support writing isochronous data to the
10 disk 64a or to support reading isochronous data from the disk 64a. In yet another embodiment,
11 the reservation facility 48E limits movement of the head 64b with respect to the disk 64a so as to
12 constrain the head 64b to a predetermined region of the disk 64a, thereby reserving a resource
13 within the switched node 62.

14 In the embodiment shown in FIG. 2A, the switched nodes constituting the multi-
15 dimensional computer network 6 may or may not comprise a disk 64a and a head 64b as shown
16 in FIG. 6. In addition, in one embodiment a select number of the switched nodes comprise
17 adapter circuitry for connecting to an external entity (e.g., a client computer). In yet another
18 embodiment, a select number of the switched nodes comprise a microprocessor for implementing
19 a distributed routing algorithm.